AD-A033 311 TEXAS UNIV AT AUSTIN DEPT OF ELECTRICAL ENGINEERING F/G 12/1
THE STRONG UNIFORM CONSISTENCY OF NEAREST NEIGHBOR DENSITY ESTI--ETC(U)
OCT 76 L P DEVROYE, T J WAGNER AF-AFOSR-2371-72

UNCLASSIFIED AFOSR-TR-76-1216 NL

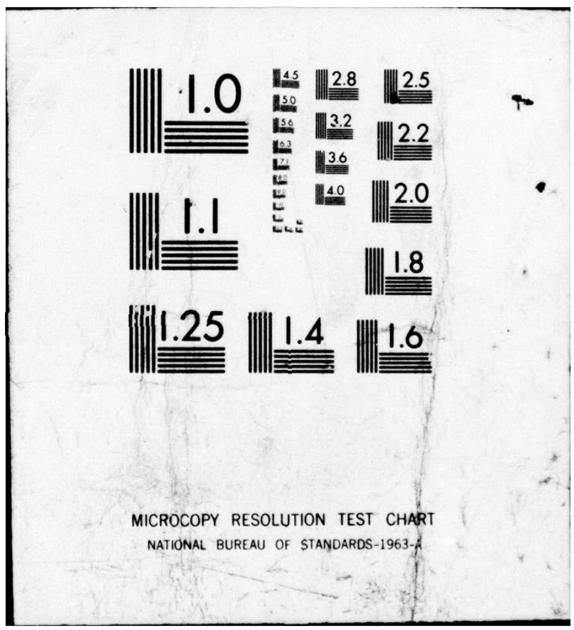
1 OF 1

AD A033311

END

DATE FILMED

1-77


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AFOSR - TR - 76 - 1216

THE STRONG UNIFORM CONSISTENCY OF NEAREST NEIGHBOR DENSITY ESTIMATES

Running Head: NEAREST NEIGHBOR DENSITY ESTIMATES

Luc P. Devroye and T.J. Wagner*

* Research of the authors was sponsored by AFOSR GRANT 72-2371.
AMS 1970 subject classifications. 60F15, 62G05.
Key Words and Phrases. Nonparametric density estimation, multivariate density estimation, uniform consistency, consistency.

THE STRONG UNIFORM CONSISTENCY OF

NEAREST NEIGHBOR DENSITY ESTIMATES

Summary. Let $X_1, \ldots, X_n$ be independent, identically distributed random vectors with values in $\mathbb{R}^d$ and with a common probability density f. If $V_k(x)$ is the volume of the smallest sphere centered at x and containing at least k of the $X_1, \ldots, X_n$ then $f_n(x) = k/(nV_k(x))$ is a nearest neighbor density estimate of f. We show that if $k = k(n)$ satisfies $k(n)/n \to 0$ and $k(n)/\log n \to \infty$ then $\sup_x |f_n(x) - f(x)| \to 0$ w.p.1 when f is uniformly continuous on $\mathbb{R}^d$.

Introduction.

Suppose that $X_1, \ldots, X_n$ are independent, identically distributed random vectors with values in $\mathbb{R}^d$ and with a common probability density f. If $V_k(x)$ is the volume of the smallest sphere centered at x and containing at least k of the random vectors $X_1, \ldots, X_n$, then Loftsgaarden and Quesenberry (1965), to estimate $f(x)$ from $X_1, \ldots, X_n$, let

$$f_n(x) = k/(nV_k(x)) \tag{1}$$

where $k = k(n)$ is a sequence of positive integers satisfying

$$\begin{aligned} &\text{(a)} \quad k(n) \uparrow \infty \\ &\text{(b)} \quad k(n)/n \to 0 \ . \end{aligned} \tag{2}$$

(The factor $k-1$ was used instead of k by Loftsgaarden and Quesenberry; this has no effect on any of the asymptotic results stated here.) They showed that $f_n(x)$ is a consistent estimate of $f(x)$ at each point where f is continuous and positive. This result can also easily be inferred from the work of Fix and Hodges (1951). For $d = 1$, Moore and Henrichon (1969) showed that

$$\sup_x |f_n(x) - f(x)| \to 0 \text{ in probability}$$

if f is uniformly continuous and positive on $\mathbb{R}$ and if, additionally,

$$k(n)/\log n \to \infty \ . \tag{3}$$

Wagner (1973) showed that $f_n(x)$ is a strongly consistent estimate of $f(x)$ at each continuity point of f if, in addition to (2b),

$$\sum_1^{\infty} e^{-\alpha k(n)} < \infty \text{ for all } \alpha > 0 \ . \tag{4}$$

(Notice that (4) is always implied by (3) but (2a) and (3) are needed to imply (4).) The result of this paper is the following theorem.

Theorem. If f is uniformly continuous on $\mathbb{R}^d$ and if k(n) satisfies (2b) and (3) then

$$\sup_x |f_n(x) - f(x)| \overset{n}{\to} 0 \text{ w.p.1.}$$

If

$$\hat{f}_n(x) = \sum_{i=1}^{n} K((x-X_i)/r(n))/nr(n)^d ,$$

where K is the uniform probability density for the unit sphere in $\mathbb{R}^d$ and $\{r(n)\}$ is a sequence of positive numbers, the recent results of Moore and Yackel (1977) (see Theorem 3.1) and the above theorem immediately yield that

$$\sup_x |\hat{f}_n(x) - f(x)| \to 0 \text{ w.p.1}$$

whenever f is uniformly continuous on $\mathbb{R}^d$ and $r(n) \to 0$, $nr(n)^d/\log n \to \infty$. This fact, an improvement over the previously published convergence results for the kernel estimate with a uniform kernel (e.g., see Theorem 2.1 of Moore and Yackel (1977)), also is a special case of Theorem 4.9 of Devroye (1976) who proves the same statement for all kernels K which are bounded probability densities with compact support and whose discontinuity points have a closure with Lebesgue measure 0.

Proof.

To simplify notation we assume below that multiplications are always carried out before division. Let $\epsilon > 0$ and choose $\delta > 0$ such that

$$|f(y) - f(x)| < \epsilon/2$$

whenever x and y are within a sphere of volume δ. Deferring measurability arguments for the moment,

$$P\{\sup_x |f_n(x) - f(x)| > \epsilon\} =$$

$$P\{\bigcup_x [V_k(x) < k/n(f(x) + \epsilon)]\} +$$

$$P\{\bigcup_{x: f(x) > \epsilon} [V_k(x) > k/n(f(x) - \epsilon)]\} .$$

The event $\bigcup_x [V_k(x) < k/n(f(x) + \epsilon)]$ implies that, for some x, there must be a sphere centered at x with volume less than $k/n(f(x) + \epsilon)$ and containing k of the random vectors $X_1, \ldots, X_n$. If $k/n\epsilon < \delta$ then the probability measure of such a sphere must be less than $\frac{k(f(x) + \epsilon/2)}{n(f(x) + \epsilon)}$ so that, for one of these spheres S,

$$\mu_n(S) - \mu(S) > \frac{k}{n} - \frac{k(f(x) + \epsilon/2)}{n(f(x) + \epsilon)}$$

$$= \frac{k\epsilon}{2n(f(x) + \epsilon)} \geq \frac{k\epsilon}{2n(F + \epsilon)}$$

where F is the maximum of f on $\mathbb{R}^d$, μ is the measure on the Borel subsets of $\mathbb{R}^d$ corresponding to f and μ_n is the empirical measure on the Borel subsets of $\mathbb{R}^d$ for $X_1, \ldots, X_n$. Thus, for $k/n\epsilon < \delta$,

$$P\{\bigcup_x [V_k(x) < k/n(f(x) + \epsilon)]\} \leq$$

$$P\{\sup_{S \in \mathcal{G}_n} |\mu_n(S) - \mu(S)| > k\epsilon/2n(F + \epsilon)\} \tag{5}$$

where $\mathcal{G}_n$ is the class of all spheres in $\mathbb{R}^d$ whose volume is less than $4k/n\epsilon$. Next, with $4k/n\epsilon < \delta$,

$$\bigcup_{x:f(x)>\epsilon} [V_k(x) > k/n(f(x)-\epsilon)] \subseteq$$

$$\bigcup_{x:f(x)>\epsilon} [V_k(x) > k/n(f(x) - (3\epsilon/4))]$$

which implies that, for some x with $f(x) > \epsilon$, there is a sphere S centered at x, with volume $\leq 4k/n\epsilon$, and

$$\mu(S) \geq k(f(x) - \epsilon/2)/n(f(x) - (3/4)\epsilon),$$

$$\mu_n(S) \leq k/n, \text{ and}$$

$$\mu(S) - \mu_n(S) \geq k\epsilon/4n(f(x) - (3/4)\epsilon) .$$

Thus

$$P\{ \bigcup_{x:f(x)>\epsilon} [V_k(x) > k/n(f(x) - \epsilon)] \} \leq \tag{6}$$
$$P\{ \sup_{S\in\mathcal{G}_n} |\mu(S) - \mu_n(S)| \geq k\epsilon/4nF \} ,$$

so that

$$P\{\sup_x |f_n(x) - f(x)| \geq \epsilon\} \leq 2P\{ \sup_{S\in\mathcal{G}_n} |\mu_n(S) - \mu(S)| \geq k\epsilon/4n(F+\epsilon) \} .$$

The proof will be completed if we show that for each $\epsilon > 0$

$$\sum_n P\{ \sup_{S\in\mathcal{G}_n} |\mu_n(S) - \mu(S)| \geq k\epsilon/4n(F+\epsilon) \} < \infty . \tag{7}$$

To prove (7) we employ a variation of the argument used by Vapnik and Chervonenkis (1971). In this variation use will be made of the following result. If $Y_1, \ldots, Y_n$ represent independent drawings without replacement from a population of k 0's and 1's then, for $\epsilon > 0$ and $k \geq n$,

$$P\left[\left|\left(\sum_{1}^{n} Y_i\right)/n - \mu\right| \geq \epsilon\right] \leq 2e^{-n\epsilon^2/(2\mu + \epsilon)} \tag{8}$$

where μ, the {number of 1's}/k, is assumed to be $\leq \frac{1}{2}$. Additionally (8) holds when $Y_1, \ldots, Y_n$ are Bernoulli random variables with parameter $\mu \leq \frac{1}{2}$. (Use the two-sided version of Theorem 3 of Hoeffding (1963) along with $\mu \leq \frac{1}{2}$ and $\log(1 + (\epsilon/\mu)) \geq 2\epsilon/(2\mu + \epsilon)$. See also section 6 of this paper.)

Now, if $\sup_{\mathcal{G}} \mu(\mathcal{G}) \leq M$ and $n \geq 8M/\delta^2$, an easy modification of Lemma 1 of Vapnik and Chervonenkis (1971) yields

$$\begin{aligned} &P[\sup_{\mathcal{G}} |\mu_n(A) - \mu(A)| \geq \delta] \leq \\ &2P[\sup_{\mathcal{G}} |\mu_n(A) - \mu_n'(A)| \geq \delta/2] \end{aligned} \tag{9}$$

where $\mu_n'(A)$ is the empirical measure for A with $X_{n+1}, \ldots, X_{2n}$ and $\mathcal{G}$ is any class of Borel sets in $\mathbb{R}^d$ for which

$$\sup_{\mathcal{G}} |\mu_n(A) - \mu(A)| \quad \text{and}$$

$$\sup_{\mathcal{G}} |\mu_n(A) - \mu_n'(A)|$$

are random variables. Putting $\mathcal{G} = \mathcal{G}_n$ we see that M can be taken to be $4kF/n\epsilon$. Since, for $\alpha > 0$,

$$\begin{aligned} &P[\sup_{\mathcal{G}_n} |\mu_n(A) - \mu'_n(A)| \geq \delta/2] \leq \\ &P[\sup_{\mathcal{G}_n} |\mu_n(A) - \mu_n'(A)| \geq \delta/2 \ ; \ \sup_{\mathcal{G}_n} \mu_{2n}(A) \leq \alpha M] \\ &+ P[\sup_{\mathcal{G}_n} \mu_{2n}(A) > \alpha M] \end{aligned} \tag{10}$$

we see, using (3) and putting $\delta = k\epsilon/4n(F+\epsilon)$, that (7) follows whenever both terms of the right-hand side of (10) are summable for some $\alpha > 0$. Looking at the first term, we note that it equals

$$\int_{\mathbb{R}^{2nd}} \frac{1}{(2n)!} \sum I_{[\sup_{\mathcal{G}_n} |\mu_n(A) - \mu_n'(A)| \geq \delta/2]} I_{[\sup_{\mathcal{G}_n} \mu_{2n}(A) \leq \alpha M]} dQ$$

where I_E is the indicator of the set $E \subseteq \mathbb{R}^d$ and Q is the probability measure on $\mathbb{R}^{2nd}$ for $X_1, \ldots, X_{2n}$ and where the inner summation is taken over all (2n)! permutations of $x_1, \ldots, x_{2n}$. But this last integral equals

$$\int_{\mathbb{R}^{2nd}} \frac{1}{(2n)!} \sum I_{[\sup_{\mathcal{G}_n} \mu_{2n}(A) \leq \alpha M]} \sup_{\mathcal{G}_n} I_{[|\mu_n(A) - \mu_n'(A)| \geq \delta/2]} dQ$$

$$= \int_{\mathbb{R}^{2nd}} \frac{1}{(2n)!} \sum I_{[\sup_{\mathcal{G}_n} \mu_{2n}(A) \leq \alpha M]} \sup_{\mathcal{G}'} I_{[|\mu_n(A) - \mu_n'(A)| \geq \delta/2]} dQ$$

$$\leq \int_{\mathbb{R}^{2nd}} \sum_{A \in \mathcal{G}'} I_{[\sup_{\mathcal{G}_n} \mu_{2n}(A) \leq \alpha M]} \left\{ \frac{1}{(2n)!} \sum I_{[|\mu_n(A) - \mu_{2n}(A)| \geq \delta/4]} \right\} dQ$$

where $\mathcal{G}' = \mathcal{G}'(x_1, \ldots, x_{2n})$ is any finite subclass of $\mathcal{G}_n$ which yields the same class of intersections with $\{x_1, \ldots, x_{2n}\}$ and where the inner summation is again taken over the (2n)! permutations of $x_1, \ldots, x_{2n}$. The quantity within $\{\cdot\}$ is bounded above, using (8), by

$$2e^{-n\delta^2/(32\mu_{2n}(A) + 4\delta)}$$

whenever $\mu_{2n}(A) \leq \frac{1}{2}$. Since $M = 4kF/n\epsilon$ we see, from (3), that for all n sufficiently large the last integral is upper-bounded by

$$2\int_{\mathbb{R}^{2nd}} e^{-n\delta^2/(32\alpha M+4\delta)}\left(\sum_{A\in G'} 1\right) dQ\ .$$

Choosing G' to be a smallest possible subclass, we have (Vapnik and Chervonenkis (1971), Cover (1965)) that $\left(\sum_{A\in G'} 1\right) \leq 1 + (2n)^{d+3}$ and, using (3) again, that the first term of (10) is summable for all $\alpha > 0$.

Looking at the second term of (10), let r be the radius of a sphere in $\mathbb{R}^d$ whose volume is $4k/n\epsilon$. If some sphere of radius r contains ℓ of the points $X_1, \ldots, X_{2n}$ then there must be at least one sphere of radius 2r, centered at one of the points $X_1, \ldots, X_{2n}$, which contains at least ℓ points. Thus

$$P[\sup_{G_n} \mu_{2n}(A) > \alpha M] \leq 2nP[\mu_{2n}(S_{X_1}(2r)) > \alpha M]$$

where $S_x(t)$ denotes the sphere of radius t centered at x. But

$$P[\mu_{2n}(S_{X_1}(2r)) > \alpha M] \leq$$

$$\max_{x\in\mathbb{R}^d} P[\mu_{2n-1}(S_x(2r)) > (\alpha 2nM-1)/(2n-1)]$$

$$\leq \max_{x\in\mathbb{R}^d} P[\mu_{2n-1}(S_x(2r)) > [(\alpha 2nM-1)/(2n-1)] - 2^d 4kF/n\epsilon]\ .$$

At this point it is not difficult, using (3) and (8), to show that the second term of (9) is summable as long as $\alpha > 2^d$.

Finally, to complete the proof, it is easy to see that all of the uncountable unions over x are indeed events and that the various supremums over G_n are indeed random variables.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE 19		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 18 AFOSR - TR - 76 - 1216	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER 9
4. TITLE (and Subtitle) 6 THE STRONG UNIFORM CONSISTENCY OF NEAREST NEIGHBOR DENSITY ESTIMATES.		5. TYPE OF REPORT & PERIOD COVERED Interim rept.,
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) 10 Luc P./Devroye and T. J./Wagner		8. CONTRACT OR GRANT NUMBER(s) 15 AF-AFOSR-2371-72
9. PERFORMING ORGANIZATION NAME AND ADDRESS University of Texas Department of Electrical Engineering Austin, Texas 78712		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F 17 A5 16 2304/A5
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research/NM Bolling AFB, Washington, D.C. 20332		12. REPORT DATE 11 October, 1976
		13. NUMBER OF PAGES 10
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) 12 12p.		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

density estimation, nonparametric density estimation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Let $X_1, \ldots, X_n$ be independent, identically distributed random vectors with values in $\mathbb{R}^d$ and with a common probability density f. If $V_k(x)$ is the volume of the smallest sphere centered at x and containing at least k of the $X_1, \ldots, X_n$ then $f_n(x) = k/(nV_k(x))$ is a nearest neighbor density estimate of f.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED 401 997 LB

20. Abstract (continued)

We show that if $k = k(n)$ satisfies $k(n)/n \to 0$ and $k(n)/\log n \to \infty$ then $\sup_x |f_n(x) - f(x)| \to 0$ w.p.1 when f is uniformly continuous on $\mathbb{R}^d$.

REFERENCES

Cover, T. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE Trans. Electronic Computers EC-10, 326-334.

Devroye, L.P. (1976). Nonparametric discrimination and density estimation. Ph.D. Thesis, University of Texas, Austin, Texas.

Fix, E., and J.L. Hodges. (1951). Discriminatory analysis. Nonparametric discrimination: consistency properties. Report 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, 21 pages.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. J. Amer. Statist. Assoc. 58, 13-30.

Loftsgaarden, D.O., and C.P. Quesenberry. (1965). A nonparametric estimate of a multivariate density function. Ann. Math. Statist. 36, 1049-1051.

Moore, D.S., and E.G. Henrichon. (1969). Uniform consistency of some estimates of a density function. Ann Math. Statist. 40, 1499-1502.

Moore, D.S., and J.W. Yackel. (to appear January 1977). Consistency properties of nearest neighbor density estimates. Ann. Statist.

Vapnik, V.N., and A. Ya. Chervonenkis. (1971). On the uniform convergence of relative frequencies of events to their probabilities. Theory Prob. Appl. 16, 264-280.

Wagner, T.J. (1973). Strong consistency of a nonparametric estimate of a density function. IEEE Trans. Systems Man Cybernetic SMC-3, 289-290.